植物分类学报
第20卷 第4期 ACTA PHYTOTAXONOMICA SINICA 1982年11月

浅谈分类学的数学方法

徐克学
(山西省晋东南地区药品检验所)

多少世纪以来,人们曾试图用数学方法解决分类问题,但是进展缓慢,直到本世纪40年代,电子计算机技术发明以后才有所突破。一经突破就得到发展和广泛应用。

30年代美国心理学家 R. C. Tryon[1] 首创聚类分析方法研究心理学。50年代电子计算机技术兴起以后,许多生物学家使用数学方法去解决分类问题。数学与生物分类学相互渗透产生了一门崭新的边缘学科——数量分类学(Numerical Taxonomy)[2,3]。数量分类学的产生,为生物分类提供了一种比较科学的方法,给生物分类学开拓出新的发展前景,古老的生物分类学正在从定性的、描述性的水平向定量的、更精确的高水平攀登,使人类对有机体的亲缘关系的认识更接近客观实际。

本文对分类的各种数学方法给予综述性介绍,并对经常使用的距离系统分类法举实例进行演算,以供有关学者参考。

数量分类学的广泛应用,促使它的数学理论迅速发展,许多数学家被吸引从事分类问题的研究,各种数学工具包括集合论、图论、概率论、信息论、统计数学和线性代数都被引用进来,应用不同的数学理论产生了不同的分类方法。现代数学最新的成果,模糊数学也被用于分类产生了模糊分类法。方法的多样性满足各种生物分类问题的不同需要。下面分别介绍:

系统分类法(Hierarchic methods of classification)这是由几何、代数和统计等运算组成的多种分类方法。60年代末期 Lance、Williams[4] 和 Wishart[5] 把六种不同的方法总结于统一的公式:

$$D_{ir}^2 = \alpha_p D_{ip}^2 + \alpha_q D_{iq}^2 + \beta D_{pq}^2 + \nu |D_{ip}^2 - D_{iq}^2|,$$

其中 D_{ip}, D_{iq} 和 D_{pq} 表示聚合前类群之间的距离, D_{ir} 表示聚合后的距离; α_p、α_q、β 和 ν 是待定参数。p 和 q 两个类群合并以后,需要计算新类群的距离系数 D_{ir},不同的一组参数给出不同的计算公式,由此获得不同的分类方法。现在已经有八种方法总结在这个公式中,见表1。这个公式如果将平方都取消,也适合于非距离系数。

表中 n_i、n_r、n_p 和 n_q 分别表示类群 G_i、G_r、G_p 和 G_q 中的分类单位个数。G_p 与 G_q 合并以后得新类群 G_r,因此 $n_r = n_p + n_q$.

这样的总结意义很大,许多不同的分类方法可以编在同一个电子计算机程序中,为分类运算工作提供方便。

系统分类法发展较早,理论和方法都比较完善,是一个比较定型的成熟的方法,它在生物分类中的应用非常广泛。

图论分类法(Graph theoretical methods of classification)组合数学中的图论应用于分类产生了图论分类法。这种方法利用无向图理论中最小生成树(Minimal spanning tree)

表 1 距离系数系统分类法参数表

方 法		α_p	α_q	β	ν
最短距离法	单联法	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{-1}{2}$
最长距离法	全联法	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
中间距离法	WPGMA 法 $(\beta = 0)$ 中线法 $\left(\beta = \frac{-1}{4}\right)$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{-1}{4} \leqslant \beta \leqslant 0$	0
	离差平方和法	$\frac{n_i + n_p}{n_i + n_r}$	$\frac{n_i + n_q}{n_i + n_r}$	$\frac{-n_i}{n_i + n_r}$	0
	重 心 法	$\frac{n_p}{n_r}$	$\frac{n_q}{n_r}$	$\frac{-n_p n_q}{n_r^2}$	0
类平均法	UPGMA 法	$\frac{n_p}{n_r}$	$\frac{n_q}{n_r}$	0	0
	可变类平均法	$\frac{(1-\beta)n_p}{n_r}$	$\frac{(1-\beta)n_q}{n_r}$	$\beta<1$	0
	可 变 法	$\frac{1-\beta}{2}$	$\frac{1-\beta}{2}$	$\beta<1$	0

的概念，把所有被分类的单位都连接在一起（图 1），再按一定的规则断开，分为二个类群。这就是图论分类法的思想。如何构造最小生成树是这个方法的关键。Prim[6] 和 Kruskal[7] 各自给出了二种不同方法去构造最小生成树。1971 年 Zahn[8] 又对图论方法作了总结。

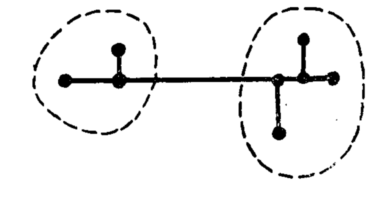


图 1

图论方法的另一个发展方向是与分支性谱系的分类(Cladistic classification)相结合，70 年代随着分子生物学的发展，图论分类法在分子遗传和分子进化中被应用于研究各种蛋白质和核酸的分类，从分子的水平上阐明遗传与进化的规律。

图论分类法的理论很不完整，有待解决的问题很多，由于应用较广，正在发展中。

主分量分类法（Principal component methods of classification） 多元统计数学中主分量分析（Principal component analysis）理论也应用于分类。在分类问题中，众多的特性都具有相关性，如果在以特性为坐标的多维空间中能找到一个方向，特性在这个方向上反映的离差变化最大，就确定了一个向量称为第一主分量，其次为第二、三、……主分量。对主分量的寻找，犹如从复杂特性的事物中抓住了主要矛盾。主分量分类法就是利用抓主要矛盾的运算技巧，把一个复杂的分类问题简化为低维空间上的简单问题，从而使分类问题迎刃而解。

主分量分类法发展较早，它的数学理论建立在线性代数的矩阵与二次型理论之上，具有严谨的数学推导，在应用时又与图论分类法相结合，更能发挥其优越性，因此它比系统分类法更完善、更细致。

主分量分析方法不仅可以解决分类问题，还可以分析生物分类系统与生活环境的关系。在生物学中具有广泛的用途。

信息分类法（Informational methods of classification） 从电信技术中发展起来的数学

理论——信息论，也被用于分类。信息论中熵或信息量，这个概念可以来描述生物类群的离散性（Diversity）。较好的分类希望得到离散程度较小，倾向于清一色的类群，这就是信息分类法的基本思想。

离散性的计算公式很多，常见的有[9]：

$$H = N\log N - \sum_{i=1}^{s} n_i \log n_i,$$

$$H = SN\log N - \sum_{i=1}^{s} [a_i \log a_i + (N - a_i)\log(N - a_i)]$$

$$H = \log N! - \sum_{i=1}^{s} \log n_i!.$$

信息论的内容虽然很丰富，但毕竟是从电信技术发展起来的，应用于生物学受到很大的限制。应生物分类的需要，新的生物信息理论正在形成和发展中。最近 Laxton[10] 做出了贡献。

信息分类的意义很大，在一定程度上它可以克服定量分类中难于解决的无序多态特性编码问题。在分子生物学中，蛋白质序列氨基酸的排列和核酸分子中核苷酸的排列都与电信编码有着类似之处，从分子水平探索生物演化的系统关系，信息分类法将有更广阔的前途。

模糊分类法（Fuzzy methods of classification） 这是当前最年青的一个分类方法，这个方法基于模糊集合概念。所谓模糊集合其实是传统集合概念的扩充。譬如在我们研究的某一高等植物类群中，把草本植物归于集合 A，数学上可以用特征函数来描述

$$f_A(x) = \begin{cases} 1 & \text{当植物 } x \text{ 属于草本} \\ 0 & \text{不属于草本。} \end{cases}$$

传统集合概念的特征函数取值非 0 即 1，也就是说植物非草本即木本，集合 A 的概念是界限分明的。可是当类群中出现草本与木本的过渡类型时，该如何处理呢？与非生命科学不同，在生物学中有机物表现的性状许多都是界限不明确的、模糊的。为了描述这种模糊的现象，需要把集合的特征函数取值加以扩充，假如特征函数值可以取到介于 0 与 1 之间的值，如此扩充了的集合概念就是模糊集。模糊集合的概念可以对生物学中那些界限不明确的模糊事物给予描述。

建立在模糊集合概念之上的分类方法称为模糊分类法。模糊分类更能适合生物分类的需要，因此它的理论发展很快。从 1965 年 Zadeh[11] 创立模糊数学以来，短短的十多年发表的论文已在 40 多篇以上，以 J. Bezdek[12,13,14] 的贡献最大，近几年又有将各种分类方法与模糊理论综合在一起的新动向。由于它的方法新颖、适合生物学的需要，很可能给定量的分类技术带来新的突破。

分类的数学方法很多，上面介绍的分类方法中，以距离系数的系统分类法应用最广。下面就这个方法举出一个演算的实例。

取桔梗科中六个种为演算的例子，特性编码数据见表 2。分类取用了 8 个特性：株高、茎缠绕与否、叶的着生方式、叶缘锯齿性状、花序、子房室数、果实开裂方式和种子是否具翼等，为了使演算尽量简单而便于说明，对特性的选取和编码做得十分粗糙，当然在实

际工作中应该做得远比此例更细致。

表 2 原始数据

编号	种名	特性 1	2	3	4	5	6	7	8
1	*Codonopsis lanceolata* Benth. et Hook. f. 羊乳	1	1	1	0	0	1	2	1
2	*C. pilosula* (Franch.) Nannf. 党参	1	1	1	0	0	1	2	0
3	*Platycodon grandiflorus* (Jacq.) A. DC. 桔梗	0	0	0	1	0	2	1	0
4	*Adenophora pereskiifolia* (Fisch.) G. Don 轮叶沙参	0	0	2	1	2	0	0	0
5	*A. remotiflora* Miq. 荠苨	0	0	0	2	1	0	0	0
6	*A. polyantha* Nakai 石沙参	0	0	0	1	2	0	0	0
	平均值	0.333	0.333	0.667	0.833	0.833	0.667	0.833	0.167
	标准差	0.516	0.516	0.816	0.753	0.983	0.816	0.983	0.408

演算的第一步将原始数据标准化。为此，先计算每个特性的平均值和标准差。若某一特性的六个数据值是 $y_i(i=1, 2,\cdots, 6)$，则

平均值 $$\bar{y}=\frac{1}{6}(y_1+y_2+\cdots+y_6),$$

标准差 $$s=\left\{\frac{1}{6-1}[(y_1-\bar{y})^2+(y_2-\bar{y})^2+\cdots+(y_6-\bar{y})^2]\right\}^{\frac{1}{2}}。$$

再连同原始数据一起代入标准化变换公式：

$$x_i=\frac{y_i-\bar{y}}{s} \qquad (i=1, 2,\cdots, 6)。$$

对每个特性都施行上面的运算，得标准化数值矩阵：

$$\begin{bmatrix} 1.291 & 1.291 & 0.408 & -1.107 & -0.848 & 0.408 & 1.187 & 2.041 \\ 1.291 & 1.291 & 0.408 & -1.107 & -0.848 & 0.408 & 1.187 & -0.408 \\ -0.645 & -0.645 & -0.816 & 0.221 & -0.848 & 1.633 & 0.170 & -0.408 \\ -0.645 & -0.645 & 1.633 & 0.221 & 1.187 & -0.816 & -0.848 & -0.408 \\ -0.645 & -0.645 & -0.816 & 1.550 & 0.170 & -0.816 & -0.848 & -0.408 \\ -0.645 & -0.645 & -0.816 & 0.221 & 1.187 & -0.816 & -0.848 & -0.408 \end{bmatrix}$$

第二步计算相似性系数。如果采用平均欧氏距离，第 i 和第 j 两个种之间的距离系数计算如下：

$$D_{ij}=\frac{1}{6}[(x_{i1}-x_{j1})^2+\cdots+(x_{i8}-x_{j8})^2]^{\frac{1}{2}} \qquad \begin{pmatrix} i=1, 2,\cdots, 6 \\ j=1, 2,\cdots, 6 \end{pmatrix},$$

其中 x_{ik} 和 x_{jk} $(k=1, 2,\cdots, 8)$分别表示第 i 和第 j 个种的标准化数据。将六个种每一对距离系数计算出来得距离矩阵 $M(0)$ (见表 3)。

第三步进行分类运算，分类运算的循环过程见表 3 。执行第一次循环时先从 $M(0)$ 中找出最小值，$D_{56}=0.592$，表明种 5 和种 6 相似性距离最近，应先将它们合并成一个新类群。新类群的距离系数需要重新计算，从表 1 给出了八种不同的计算公式，不同的计算方法得出不同的分类结果。在此例取最容易计算的最短距离法，将数值代入公式，实际上

表 3 分类运算过程

	1	2	3	4	5	6
1	0					
2	0.866	0				
3	1.553	1.289	0			
4	1.821	1.602	1.465	0		
5	1.895	1.686	1.109	1.049	0	
6	1.821	1.602	1.182	0.866	0.592	0

$M(0)$

	1	2	3	4	7
1	0				
2	0.866	0			
3	1.553	1.289	0		
4	1.821	1.602	1.465	0	
7	1.821	1.602	1.109	0.866	0

$M(1)$ $D_{56} = 0.592$, $G_7 = G_5 + G_6$

	8	3	4	7
8	0			
<3	1.289	0		
<4	1.602	1.465	0	
7	1.602	1.109	0.866	0

$M(2)$ $D_{12} = 0.866$, $G_8 = G_1 + G_2$

	8	3	9
8	0		
3	1.289	0	
9	1.602	1.109	0

$M(3)$ $D_{74} = 0.866$, $G_9 = G_7 + G_4$

	8	10
8	0	
10	1.289	0

$M(4)$ $D_{93} = 1.109$, $G_{10} = G_9 + G_3$

是取最小值运算。例如

$$
\begin{aligned}
D_{71} &= \mathrm{Min}\{D_{51},\ D_{61}\} \\
&= \mathrm{Min}\{1.895,\ 1.821\} \\
&= 1.821.
\end{aligned}
$$

计算结束后得新的矩阵 $M(1)$。

再对矩阵 $M(1)$，$M(2)$，……依次施行前面的运算，每循环一次一个类群被归并，矩阵减小一阶，直到将所有的种都归并成一个类群为止。

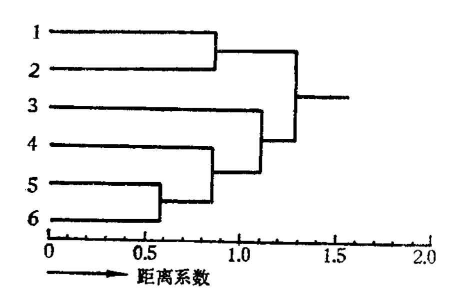


图 2 最短距离法(单联法)树系图

最后将分类结果画成树系图(图 2)。树系图不仅形象地显示出被分类单位之间的隶属关系，而且还定量地表示类群之间的结合水平。例如种 5 和种 6 在 0.592 的距离水平上相互结合。

如果将表一所提供的八种方法都算出来，现在要问究竟选取哪一个方法好呢？这个问题涉及

最优分类，最优分类是一个尚未完全解决的理论问题。这里我们给出一个比较合适的选择方法。

要问什么分类方法好，先必须确立一个判别的标准。

让我们先考虑树系图，从树系图也给出种与种之间的相似性水平，如果将每一对种之间的相似性系数都写出来，就得到一个新的矩阵称为树系图的协表矩阵（Cophenetic matrix）：

$$\begin{bmatrix} 0 & 0.866 & 1.289 & 1.289 & 1.289 & 1.289 \\ 0.866 & 0 & 1.289 & 1.289 & 1.289 & 1.289 \\ 1.289 & 1.289 & 0 & 1.109 & 1.109 & 1.109 \\ 1.289 & 1.289 & 1.109 & 0 & 0.866 & 0.866 \\ 1.289 & 1.289 & 1.109 & 0.866 & 0 & 0.592 \\ 1.289 & 1.289 & 1.109 & 0.866 & 0.592 & 0 \end{bmatrix}$$

在协表矩阵中分类结果呈现的相似性关系应该与分类之前原来的相似性关系（即矩阵 $M(0)$）尽可能一致。这个一致性显然是判断分类好坏的一个重要标准。

有了判别的标准就可以进行具体计算。二个矩阵之间的一致性有三个比较系数可以参考使用，

$$M = \max\{|D_{ij} - D_{ij}^*|\},$$

$$A = \left[\frac{2}{t(t-1)} \Sigma(D_{ij} - D_{ij}^*)^2\right]^{\frac{1}{2}},$$

和

$$R = \frac{\Sigma(D_{ij} - \bar{D})(D_{ij}^* - \bar{D}^*)}{[\Sigma(D_{ij} - \bar{D})^2 \cdot \Sigma(D_{ij}^* - \bar{D}^*)^2]^{\frac{1}{2}}}。$$

其中求最大值和求和号都是对标号 $i = 2, 3, \cdots, t$ 和 $j = 1, 2, \cdots, i-1$ 进行；D_{ij}^* 和 D_{ij} 分别表示协表矩阵和原距离矩阵的第 i 行第 j 列元素，$\bar{D}^*$ 和 $\bar{D}$ 表示其相应的平均值；t 表示矩阵的阶数。

将桔梗科的数据按表 1 所提供的各种方法进行分类运算，再对每一个分类结果算出 M，A 和 R 的值，计算结果见表 4。

表 4 分类结果的比较

方 法 （系 数）	M	A	R
最短距离法(单联法)	0.6058	0.3185	0.9085
最长距离法(全联法)	0.6058	0.2518	0.9165
WPGMA 法 ($\beta = 0$)	0.3063	0.1638	0.9122
中线法 ($\beta = -0.25$)	0.5383	0.2694	0.8950
离差平方和法	1.1115	0.5693	0.9081
重心法	0.4249	0.2105	0.9183
UPGMA 法	0.3794	0.1532	0.9190
可变类平均法 ($\beta = -0.5$)	1.6334	0.9438	0.8988
可变法 ($\beta = -0.5$)	1.1991	0.6395	0.9122

从比较中看出 UPGMA 法和 WPGMA 法二个分类结果优于其它的结果。最后画出这二个分类结果的树系图（图 3，4）。二个树系图是经过大量运算然后精心挑选出来的，

由于它们与原来的相似性关系有较高的拟合度，二个结果差异甚微。

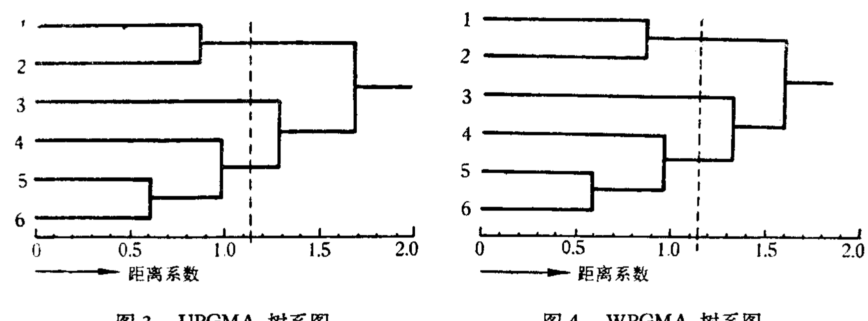


图 3 UPGMA 树系图　　图 4 WPGMA 树系图

从二张树系图清楚地看到党参与羊乳有较密切的关系，它们同属于党参属（*Codonopsis*），轮叶沙参，荠苨和石沙参三个种比较接近，它们同属于沙参属（*Adenophora*）。图中虚线表示区别属的截线。桔梗单独另立一属，桔梗属（*Platycodon*），该属与沙参属比较接近。定量分类的结果与传统分类非常吻合。它说明我们这个演算的例子尽管特性的选取和编码都十分简单，定量分类的方法仍然保持较高的可靠性。

全部数值运算由中国科学院计算中心 TQ-16 型电子计算机完成。

参 考 文 献

[1] Tryon, R. C. and D. E. Bailey, 1970: Cluster Analysis.
[2] Sokal, R. R. and P. H. A. Sneath, 1963: Principles of Numerical Taxonomy.
[3] Sneath, P. H. A. and R. R. Sokal, 1973: Numerical Taxonomy.
[4] Lance, G. N. and W. T. Williams, 1967: *Comput. J.*, **9**: 373—380.
[5] Wishart, D., 1969: *Biometrics*, **22**: 165—170.
[6] Prim, R. C., 1957: *Bell System Tech. J.*, **36**(6): 1389—1401.
[7] Kruskal, J. B., 1956: *Proc. Amer. Math. Sco.*, **7**: 48—50.
[8] Zahn, C. T., 1971: *IEEE Trans. Comput.*, **C-18**: 68—86.
[9] Clifford, H. T. and W. Stephenson, 1975: An Introduction to Numerical Classification.
[10] Laxton, R. R., 1978: *J. Theor. Biol.*, **70**(1): 51—67.
[11] Zadeh, L. A., 1965: *Inform. Contr.*, **8**: 338—353.
[12] Bezdek, J., 1973: *J. Math. Biol.*, **1**(1): 57—71.
[13] Bezdek, J., 1974: *J. Cybern.* **3**(3): 58—71.
[14] Bezdek, J. and J. D. Harris, 1978: *Fuzzy set and Systems*, **1**(2): 111—127.

A PRELIMINARY INTRODUCTION TO MATHEMATICAL METHODS FOR TAXONOMY

Xu Ke-xue

(Laboratory for the control of drugs in Jindongnan locality Shanxi province)

Abstract

In this paper, the various mathematical methods applied to taxonomy are introduced to readers. Some approaches to the classification induced by statistics, graph theory, information theory, fuzzy mathematics are discussed. An example of classification (6 OTU's with 8 characters) is given for convenience of discussion. The original data matrix of this example is obtained from 6 species in the family of *Campanulaceae*.